CARE: Extracting Experimental Findings From Clinical Literature

Aakanksha Naik[1] **Bailey Kuehl**[1] **Erin Bransom**[1]
Doug Downey[1] **Tom Hope**[1,2]

[1]Allen Institute of AI, USA
[2]The Hebrew University of Jerusalem

Abstract

Extracting fine-grained experimental findings from literature can provide dramatic utility for scientific applications. Prior work has developed annotation schemas and datasets for limited aspects of this problem, failing to capture the real-world complexity and nuance required. Focusing on biomedicine, this work presents CARE—a new IE dataset for the task of extracting clinical findings. We develop a new annotation schema capturing fine-grained findings as n-ary relations between entities and attributes, which unifies phenomena challenging for current IE systems such as discontinuous entity spans, nested relations, variable arity n-ary relations and numeric results in a single schema. We collect extensive annotations for 700 abstracts from two sources: clinical trials and case reports. We also demonstrate the generalizability of our schema to the computer science and materials science domains. We benchmark state-of-the-art IE systems on CARE, showing that even models such as GPT4 struggle. We release our resources[1] to advance research on extracting and aggregating literature findings.

1 Introduction

> It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials. (Archie Cochrane, 1979)

Though this critique focused on clinical trials, the statement arguably applies to much of science today. There is tremendous potential utility in extracting, structuring and aggregating fine-grained information about experimental findings and the conditions under which they were achieved, across scientific studies. Once extracted and aggregated, scientific findings can power many critical applications such as producing literature reviews (DeYoung et al., 2021), supporting evidence-based decision-making (Naik et al., 2022), and generating new hypotheses (Wang et al., 2023).

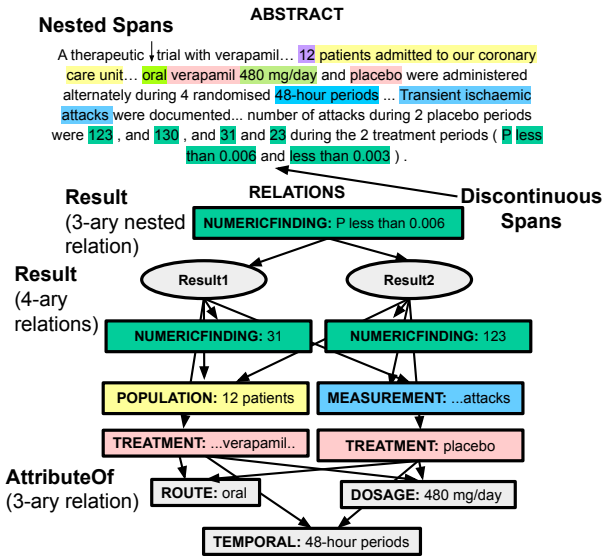

Figure 1: A partial example of entity, attribute and relation annotation using our schema for a clinical trial.

While there have been efforts on building resources and tools to capture findings in various domains such as clinical trials (Lehman et al., 2019), computer science (Jain et al., 2020) and social and behavioral sciences (Magnusson and Friedman, 2021)—a major obstacle has been creating a representation that is *expressive* enough to capture complex and nuanced information about findings. We propose a new representation schema that makes important progress in capturing the real-world complexity of scientific findings in papers, and use it to build a high-quality annotated dataset focusing on biomedical (clinical) findings. Our schema represents fine-grained information about experimental findings and conditions as n-ary relations between entities and attributes, and includes several structural complexities such as discontinu-

[1]CARE is available at https://github.com/allenai/CARE

ous span annotation, variable arity in relations and nestedness in relations. These aspects have been studied individually in previous datasets (Karimi et al., 2015; Tiktinsky et al., 2022), but our schema is the first to unify them. Our dataset also captures *numeric* findings in addition to their interpretation (e.g., significance, utility, etc.); prior datasets typically focus solely on the latter (e.g., Lehman et al. (2019) captures *increases/decreases* in outcomes but not their magnitudes).

Though these factors make our annotation schema more complex than prior work, the additional nuance it affords can power several high-impact downstream use cases. For instance, the process of conducting a systematic review[2] includes extracting data about specific outcomes from a large number of relevant studies. Unlike prior schemas which only capture interpretation and not precise numeric outcomes, extractors trained on our schema can extract relevant data to assist the review process. Moreover, fine-grained population and treatment details captured by our schema makes it capable of answering highly complex clinical questions (e.g., "Does vaccination significantly improve mortality outcomes for female patients over 65 who required ventilation?"). Therefore, models trained on our schema can assist physicians in developing more personalized treatment plans, especially for patients with multiple comorbidities or health conditions. Our schema can also help with the development and exploration of richer clinical hypotheses due to its additional granularity.

To build our dataset, named CARE (Clinical Aggregation-oriented Result Extraction), we collect extensive annotations for 700 abstracts (clinical trials and case reports). We also conduct annotation studies demonstrating that our schema generalizes to computer science and materials science, using minor updates based on analogies between aspects across experimental domains (e.g., *populations/interventions* $\rightarrow$ *tasks/methods* in CS). This reflects the expressive power of our schema to generalize across domains while capturing granular and useful information, making it a strong "backbone schema" for research efforts on result-oriented scientific IE.

We achieve good agreement scores (0.74-0.78 partial F1) comparable to prior work that used simpler schemas that are easier to annotate (Luan et al., 2018; Nye et al., 2018), and at the same time our resulting dataset is larger in size than previous corpora. Our final dataset annotation is extremely rich; at 16.23 relations per abstract, our relation density is nearly 4x that of prior work on annotating findings from clinical trials (Lehman et al., 2019).

We evaluate a wide range of IE models on our dataset, including both extractive systems and generative LLMs. Given the high annotation burden, we test generative LLMs in both fully supervised as well as zero-shot and few-shot settings. Our results demonstrate the difficulty of our dataset, with even SOTA models such as GPT4 struggling to accurately extract clinical findings. As a highly challenging new dataset designed to be reflective of real-world nuance and informational needs, we hope CARE[3] is an important resource for the scientific NLP and IE research community to pursue.

2 Related Work

2.1 Information Extraction from Scientific Literature

Much prior work has focused on information extraction from scientific papers (Luan et al., 2018; Jain et al., 2020), including biomedical literature (see (Luo et al., 2022a) for a detailed summary). Most relevant to our goal in this work is prior research on extracting findings or results from scientific literature, which has only explored limited aspects of this problem.

Gábor et al. (2018) and Luan et al. (2018) annotate *associative* relations between entities being compared or producing a result, as part of their broader goal of developing IE resources for computer science, but do not capture any nuance (e.g., directionality, causality, etc. of results). Conversely, Magnusson and Friedman (2021) develop a schema focused solely on capturing associations between experimental variables and evidence. However, their focus on sentence-level annotation from scientific claims limits how much additional nuance about experimental setting can be captured.

Some prior efforts have also explored result extraction from biomedical literature. The EBM-NLP (Nye et al., 2018) and Evidence Inference (Lehman et al., 2019) corpora contain annotations for experimental findings from clinical trials, following the well-established PICO (participant, intervention, comparator, outcome) framework (Richardson et al., 1995). Sanchez-Graillet et al. (2022) also

[2] https://training.cochrane.org/handbook/current

[3] https://github.com/allenai/CARE

Type	EBM	CTKG	Example
Population	✓	✓	This study compared rizatriptan 5 mg and placebo in 1268 outpatients treating a single migraine attack
Subpopulation	✓	✓	We found low-certainty evidence of little or no difference in delirium (RR 1.06, 95% CI 0.55 to 2.06; 2 studies, 800 participants)
Treatment	✓	✓	Dialysate magnesium was 0.375 mM/L for the hemodialysis
Measurement	✓	✓	Headache relief rates after rizatriptan 10 mg were higher
Temporal	✗	✓	After a 48-hour run-in period , oral verapamil 480 mg/day and placebo were administered
NumericFinding	✗	✓	The number of attacks during treatment periods were 31 and 23
Qualifier	✗	✗	Pindolol and metoprolol lowered blood pressure to the same extent

Table 1: Examples of entity types in our schema. EBM and CTKG columns indicate whether these entity types are present in the EBM-NLP and CTKG schemas respectively. EBM-NLP uses IE to extract information according to its schema, while CTKG is a database schema not based on IE.

develop a PICO-inspired schema-based annotation format for diabetes and glaucoma trials. Chen et al. (2022) focuses on aggregating findings, which are already manually organized in structured format in databases such as AACT (Aggregate Analysis of ClinicalTrials.gov) (Tasneem et al., 2012). However, these efforts are tailored to clinical trials and do not translate easily to other domains. Finally, Luo et al. (2022a) conducted *novelty* annotation for relations, indicating whether they were presented as new observations; however they did not focus on experimental findings.

In contrast, we develop a representation schema expressive enough to capture fine-grained experimental findings, while generalizing across scientific domains. Our schema also contains phenomena challenging for SOTA IE models (§3.2).

2.2 Extracting Numeric Information

Another unique aspect of our schema is our focus on capturing numeric information from experimental findings and setup, which is understudied. Some prior work on open IE has explored extraction and linking of numeric spans (Madaan et al., 2016; Saha et al., 2017), including linking to implied entities (Elazar and Goldberg, 2019) (e.g., "it's worth two million" can be linked to currency). However, these models broadly focused on sentence-level extraction and did not evaluate on scientific text.

Within the scientific domain, some studies have focused on numeric information extraction from biomedical/clinical text. Kang and Kayaalp (2013) and Claveau et al. (2017) extract numeric spans from FDA-released descision summaries and clinical trial eligibility criteria respectively. EBM-NLP (Nye et al., 2018) annotates some categories of numeric information associated with cohorts participating in a clinical trial, but ignores trial outcomes and findings. Among non-medical scientific domains, numeric span extraction work has mainly focused on extraction from tables (Hou et al., 2019). None of these studies focus extensively on linking numeric spans with entities that can help in interpreting this information, which is key to our work.

Type	EBM	CTKG	Example
Age	✓	✗	for those age 60-67 years
Sex	✓	✗	210 females
Size	✓	✓	12 patients
Condition	✓	✓	patients getting hemodialysis
Demographic	✗	✗	A 40's Japanese man
Route	✗	✗	oral verapamil
Dosage	✗	✗	verapamil 480 mg/day
Strength	✗	✗	rizatriptan 5 mg
Duration	✗	✗	for 4 weeks

Table 2: Examples of attribute types in our schema. EBM and CTKG columns indicate whether these types are present in the EBM-NLP and CTKG schemas.

3 Annotation Schema

We develop a new annotation schema to represent fine-grained clinical findings present in biomedical abstracts, and later demonstrate its broader applicability to domains beyond biomedicine (§6.2). Our schema captures this knowledge via three main elements, commonly used in IE tasks:

1. Entities involved in a study, which are spans of text, either contiguous or non-contiguous, belonging to one of the seven types listed in Table 1.

2. Attributes associated with entities, which are also contiguous or non-contiguous spans of text, belonging to one of the nine types listed in Table 2. The first five attribute types are associated with

Type	Arity	EI	CTKG	Example
AttributeOf	N-ary	✗	✗	($Subpopulation$: 144 had the U-type method, $Size$: 144)
SubpopulationOf	N-ary	✗	✗	($Population$: 285 women, $Subpopulation$: 144 had the U-type method, $Subpopulation$: 141 had the H-type method)
TreatmentOf	Binary	✗	✓	($Subpopulation$: 144 had the U-type method, $Treatment$: U-type method)
Result	N-ary	✓	✓	($Subpopulation$: 144 had the U-type method, $Measurement$: objective cure rates, $NumericFinding$: 87.5%)

Table 3: Examples of relation types in our schema. EI and CTKG columns indicate whether these relation types are present in the EI and CTKG schemas respectively. While the EI and CTKG datasets contain 4-ary and binary result relations respectively, our n-ary schema allows fine-grained information to be captured more flexibly.

population and subpopulation entities, while the remaining four types are associated with treatment entities. Other entity types do not have any associated attributes.

3. N-ary Relations linking together various entities and attributes, where N (relation arity) is variable and nesting is allowed. A relation is an n-tuple, where each element can be an entity, attribute or another n-ary relation. Relations are categorized into four types listed in Table 3.

3.1 Comparison to Existing Clinical Schemas

Prior work such as EBM-NLP (Nye et al., 2018) and Evidence Inference (Lehman et al., 2019; DeYoung et al., 2020) has focused on developing IE schemas to represent clinical knowledge appearing in the literature in a structured format. In addition, work such as CTKG (Chen et al., 2022) outside the NLP/IE sphere has built schema for representing clinical information in databases. However, these schemas suffer from a few shortcomings: (i) most are designed for clinical trials; their applicability to other types of biomedical literature is untested, (ii) focus on a small set of broad entity types, which leaves out fine-grained details, (iii) follow strict relation formats, which makes it hard to capture additional nuance that might be useful for interpreting findings.

Our schema makes several enhancements to tackle these issues. First, it is extensible to other categories of biomedical literature beyond clinical trials, and we demonstrate this by applying our schema to case reports. Second, our schema captures more fine-grained information about various entities than prior work via attributes (see Table 2). Third, allowing for variable arity and nesting in relation annotation provides the flexibility which makes our schema capable of representing both atomic findings (e.g., value of primary outcome observed for a given treatment) as well as composite findings (e.g., outcome improvement observed for treatment vs control groups). Tables 1, 2 and 3 provide a more detailed comparison of our schema with EBM-NLP, EI and CTKG.

3.2 Annotation Complexity

In addition to using an expanded set of entity, attribute and relation types, our annotation schema supports the following phenomena (also illustrated in Figure 1), unifying them all in a single dataset:

Discontinuous spans: Biomedical abstracts often present multiple entities as conjunctive phrases or lists of items, so we allow discontinuous span annotation to capture every entity. For example, given the phrase "maximal diameters and volumes", our scheme captures two measurement entities: "maximal diameters" and "maximal volumes", with the latter being a discontinuous span.

Nested/overlapping spans: Attributes, as defined in our annotation scheme, are often present within an entity span or overlap with an entity span. This motivates our decision to allow nested and overlapping spans to be annotated.

Variable arity in relations: Owing to variation in clinical studies, findings are often described in a wide range of formats (e.g., outcome for a single population, outcome for a pair of populations, outcome for a single population at different time periods, etc.). This diversity motivated our choice of *variable arity* for relation annotation, similar to Tiktinsky et al. (2022).

Nested relations: In addition to outcomes for individual populations/groups, clinical studies often present comparative findings and analyses, such as improvement on an outcome given a pair of interventions. Our scheme allows for annotation of nested relations to link these higher-order observations with their associated atomic findings.

Our complete annotation guidelines can be found at `https://github.com/allenai/CARE`. Figure 1 presents partial entity, attribute and relation annotations for an example clinical trial abstract.

Category	Exact F1	Partial F1
Entity	0.5764	0.7578
Attribute	0.6174	0.7801
Relation	0.4209	0.7414

Table 4: Final inter-annotator agreement scores on a sample of 28 abstracts, measured during full-scale data annotation.

Metric	Train	Dev	Test
#Docs	500	100	100
#Tokens	135,363	27,120	25,219
#Entities	12022	2367	2286
#Attributes	3992	804	762
#Relations	8205	1594	1560

Table 5: Statistics for final collected dataset.

4 Dataset Collection

Annotation Tool: We use TeamTat[4] (Islamaj et al., 2020), a web-based tool for team annotation since it allows for n-ary and nested relation annotation, a core component of our schema.

Annotator Background: We recruit two in-house annotators[5] with backgrounds in data analytics and data science, both having extensive experience in reading and annotating scientific papers. One of our annotators has a background in biology. Both annotators went through several pilot rounds to gain familiarity with our task and schema. Additionally, we used their feedback and insights from pilots to solidify our schema design (see §4.1). We also solicited feedback from two medical students and an MD to validate our final schema.

Data Sources: CARE covers two categories of biomedical literature: (i) clinical trials, and (ii) case reports. Clinical trials are research studies that test a medical, surgical, or behavioral intervention in people to determine whether a new form of treatment or prevention or a new diagnostic device is effective. Case reports are detailed reports of the symptoms, signs, diagnosis, treatment, and follow-up of an individual patient, usually motivated by unusual or novel occurrences. We sample clinical trials from the EBM-NLP (Nye et al., 2018) dataset, which consists of 4993 abstracts annotated with PICO spans, only retaining abstracts containing at least one number (4685 in total). To sample case reports, we extract all reports with at least one number in the abstract from PubMed (907,862 in total) and randomly sample from this pool. We sample 350 abstracts from each source, resulting in our final dataset size of 700 abstracts, which is slightly larger than other prior corpora that perform fine-grained annotation (§ 4.3). Further characteristics of our abstract sample are detailed in Appendix C.

[4] https://www.teamtat.org

[5] included as co-authors on this paper

4.1 Annotation Pilots

We conducted three pilot rounds with the following goals: (i) training annotators to apply our schema, (ii) evaluating agreement, and (iii) assessing whether our schema captures clinical knowledge of interest. Annotators worked on a fresh set of 5-10 abstracts per round, followed by agreement computation and disagreement discussion. For entity and attribute annotation, agreement is computed as entity-level F1 between annotators, using both strict (entity boundaries match exactly) and partial (entity boundaries overlap on at least one token) matching. For relations, we first align annotations from both annotators by linking pairs of relations which share $\geq 50\%$ of participating entities. Agreement is computed as F1 score between annotators, using both strict (100% of entities match) and partial matching. After achieving reasonable agreement levels by round 3 (partial F1 scores of 0.79, 0.68 and 0.79 for entity, attribute and relation annotation respectively), we started full-scale data annotation (further discussion in Appendix C).

4.2 Full-Scale Annotation

The full-scale data annotation process was conducted in six rounds. To continue monitoring agreement, a small agreement set of 5 abstracts (not identified to the annotators) was included in every round. Table 9 in the appendix presents inter-annotator agreement during each annotation round, while Table 4 shows overall agreement scores. Overall and per-round agreement scores continued to remain in the same range as agreement scores from later pilot rounds, demonstrating consistency in annotation quality. Despite the complexity of our schema, our agreement scores are comparable to datasets using simpler schemas like EBM-NLP (entity agreement of 0.62-0.71; Cohen's kappa) and SciERC (relation agreement of 67.8; kappa score). Appendix C provides additional details about our full-scale annotation setup.

Consensus Annotation: For all abstracts annotated by multiple annotators during pilots or full-

Phenomenon	Train	Dev	Test
#Discontinuous Spans	8.9%	10.1%	9.3%
#Nested Spans	3.4%	4.3%	2.5%
#Overlapping Spans	1.6%	2.0%	0.7%
#Nested Relations	11.4%	11.2%	11.9%

Table 6: Prevalence of interesting annotation phenomena in final collected dataset.

scale annotation (55 in total), we construct a "consensus" version post disagreement discussion. The final dataset releases consensus annotations for these abstracts. Since this subset has been annotated by multiple annotators and discussed extensively, we expect annotations to be higher-quality and include all these abstracts in the test set.

4.3 Dataset Statistics

Table 5 gives an overview of statistics for our final collected dataset. Our dataset size is comparable to other prior biomedical corpora which performs exhaustive fine-grained annotation (though not always with a clinical knowledge focus) such as BioRED (Luo et al. (2022a); 600 abstracts) and Sanchez-Graillet et al. (2022) (211 abstracts). Table 6 presents the proportion of various interesting phenomena allowed by our schema in the final dataset. Interestingly, CARE contains 9% discontinuous spans, making it one of the rare datasets containing a large proportion of discontinuous mentions.[6] At 11%, the final data also contains a high proportion of nested relations.

5 Benchmarking IE Models

We benchmark the performance of two categories of models on CARE: (i) extractive models, and (ii) generative LLMs. We also test generative LLMs in two settings: (i) finetuning on the full training set, and (ii) zero-shot and in-context learning.

Experimental Setup: We test each model on the three sub-tasks—entity extraction, attribute extraction and relation extraction—in isolation. Model performance on entity and attribute extraction is evaluated using entity-level F1. Relation extraction performance is evaluated using a relaxed overlap F1 score metric inspired by Tiktinsky et al. (2022), which assigns partial credit to correctly identified subsets of entities in a relation, even if all identified entities do not match. As with agreement score calculation, predicted relations are first aligned with gold relations by choosing the gold relation with highest overlap per predicted relation. Then a partial match score is computed as $\#shared\_entities/total\_entities$ and used in the F1 computation instead of binary 0/1 score.

5.1 Extractive IE Baselines:

We evaluate the following systems:

- **OneIE** (Lin et al., 2020): A sentence-level joint entity, relation and event extraction system, which extracts an "information network" representation of entities and events (nodes), connected by relations (edges). Beam search is used to find the highest-scoring network.
- **PURE** (Zhong and Chen, 2021): A sentence-level pipelined extraction system, which learns separate contextual representations for entity and relation extraction, using entity representations to further refine relation extraction.
- **LocLabel** (Shen et al., 2021): A sentence-level two-stage named entity recognition (NER) system capable of extracting nested spans. Inspired by object detection work, it produces boundary proposals for candidate entities, then labels them with correct entity types.
- **W2NER** (Li et al., 2022): A sentence-level unified NER model, capable of extracting nested and discontinuous spans. It recasts NER as word-word relation classification on a 2-D grid of word pairs, then decodes word pair relations into final span extractions.

For comparability and better adaptation to our dataset, we replace BERT-based encoders in all systems with PubmedBERT (Gu et al., 2021), and follow best-reported hyperparameters per system (see Appendix E). Table 7 presents their performance on entity and attribute extraction. Unfortunately, applying these systems to our relation extraction task is infeasible, since none of them are designed for document-level relation extraction or n-ary relations. Tiktinsky et al. (2022) modify PURE for n-ary relation extraction with variable arity. However, given a set of candidate entities, they consider all possible n-ary combinations and predict relationships per cluster. This is tractable for their work on sentence-level extraction of single-type (drug interaction) relations, but not tractable for document-level multi-type n-ary relation extraction.[7] Therefore, we do not test extractive models

[6] Dai et al. (2020) considers 10% discontinuous spans to be a high proportion, identifying only three biomedical datasets that satisfy this criterion: CADEC (Karimi et al., 2015), ShARe 13 (Pradhan et al., 2013) and ShARe 14 (Mowery et al., 2014).

[7] On limiting combination size to 10, every abstract pro-

Model	Ent F1	Attr F1	Rel F1
Extractive Baselines			
OneIE	55.07	48.84	–
PURE	**55.94**	**61.04**	–
LocLabel	53.69	55.25	–
W2NER	51.84	57.98	–
Generative Baselines			
FLAN-T5	**45.08**	23.27	33.24
BioGPT	14.43	**29.84**	33.15
BioMedLM	1.50	10.62	32.76
GPT-3.5 0-shot	11.14	5.06	14.35
GPT-3.5 1-shot	21.40	8.61	31.58
GPT-3.5 3-shot	23.40	8.85	31.58
GPT-3.5 5-shot	8.92	9.92	32.20
GPT-4 0-shot	26.89	9.02	32.04
GPT-4 1-shot	31.07	11.82	42.81
GPT-4 3-shot	16.68	13.16	53.69
GPT-4 5-shot	5.04	13.90	**55.04**

Table 7: Performance of all extractive and generative baselines on entity, attribute and relation extraction.

on relation extraction.

Another caveat with extractive models is that they do not identify discontinuous spans (except W2NER). To assess how this impacts model performance, we compute an additional entity-level F1 score which merges span predictions linked in gold annotation (i.e., we assume oracle span merging), and observe that this does not significantly improve performance (avg. increase of $\sim$1.5 F1). Therefore, Table 7 reports F1 scores without merging.

5.2 Generative IE Baselines:

Motivated by recent work demonstrating LLM capabilities on information extraction (Wadhwa et al., 2023), we assess the ability of LLMs on our tasks, in both finetuning and zero-shot/in-context learning settings.

We evaluate the following finetuned LLMs:

- **FLAN-T5** (Chung et al., 2022): Enhanced version of T5 (Raffel et al., 2020) finetuned on a large mixture of tasks, but not specifically pretrained for biomedicine. We use FLAN-T5-XL, which has 3B parameters.
- **BioGPT** (Luo et al., 2022b): A 1.6B autoregressive model, pretrained from scratch on 15M abstracts and titles from PubMed with a custom Pubmed-trained tokenizer.
- **BioMedLM**[8]: A 2.7B autoregressive model, pretrained from scratch on all PubMed abstracts and full-texts from the Pile (Gao et al., 2020) with a custom PubMed-trained tokenizer.

When training and testing on attribute and relation extraction, these models are provided gold entities and attributes by surrounding them with entity markers ($< ent >< /ent >$) in the input.

We evaluate GPT3.5 and GPT4 in zero-shot and in-context learning settings. We provide our IE schema and example outputs and prompt the model to produce extractions in a clean JSON format that adheres to the schema. Additionally, for our in-context learning experiments, we follow (Liu et al., 2021) and select the k *most similar* examples from the training set for every test instance according to similarity computed by the SPECTER v2.0 (Singh et al., 2022) PRX model trained on scientific titles and abstracts. Selected examples are appended to the prompt in decreasing order of similarity, with later examples dropped if they don't fit. We run experiments for the $k = 1, 3, 5$ most similar examples. Further hyperparameter details for all models are provided in Appendix E and full prompts are provided in Appendix G.

Table 7 shows the performance of all generative models. One caveat with GPT3.5/4 is that model outputs sometimes contain correct entity/attribute spans assigned to the wrong type (e.g., a subpopulation misclassified as a population entity in a result relation). Since we are evaluating the performance of relation extraction in isolation, we do not consider such mistyping as errors.

5.3 End-to-End Evaluation:

In addition to evaluating SOTA systems on each sub-task in isolation, we assess the feasibility of end-to-end extraction. Table 7 shows that PURE is the best-performing system on entity and attribute extraction. On the other hand, GPT4 5-shot and FLAN-T5 perform best on relation extraction (GPT3.5 5-shot and BioGPT are close). We test out a hybrid end-to-end extraction system in which entities and attributes are detected using PURE, then input text marked up with these extractions is provided to FLAN-T5 for relation extraction. This hybrid system achieves an F1 score of 33.58, very similar to RE performance with gold markup. Hypothesizing that this might be an indication that finetuned LLMs ignore entity/attribute markup during RE, we run an additional experiment in which we train FLAN-T5 to extract relations from raw text (no markup). This setup achieves an F1 score

duces 500,000 candidate combinations

[8] https://crfm.stanford.edu/2022/12/15/biomedlm.html

Original Type	Generalized Type	Description
Population	Research Problem Context	Setting/scenario in which the authors are testing their hypothesis (e.g., task or dataset being studied in ML/NLP).
Subpopulation	Problem Stages/Sub-parts	Subgroups or subsamples of overall setting (e.g., dataset splits in ML/NLP).
Treatment	Technique/Method	Key technique being proposed or investigated and other techniques being compared (e.g., model or metric in ML/NLP).
SubpopulationOf	Sub-PartOf	Links together problem context entities to stage/sub-part entities (e.g., for ML/NLP, this relation would link the overall task to low-data and fully supervised settings).
TreatmentOf	AppliedTo	Links together a technique to all the problem contexts/sub-parts it is being tested in.

Table 8: Changes required to construct a generalized version of our original schema developed for clinical finding extraction, which we use to test whether it applies to other domains such as computer science and materials science

of 33.07, showing that entity/attribute markup does not provide significant benefit.

6 Discussion

6.1 How much does strict evaluation underestimate LLM performance?

Table 7 shows that even fully-supervised generative models severely lag behind much smaller extractive models on entity and attribute extraction. However, prior work (Wadhwa et al., 2023) has observed that strict IE evaluation metrics underestimate the performance of LLMs since their outputs often contain minor variations from gold annotations, which could still be correct. Therefore, we conduct human evaluation of a subset of FLAN-T5 and GPT4 5-shot predictions on entity and attribute extraction for a more accurate assessment.

For every setting, we collect all abstracts with one or more wrong predictions and randomly sample ten to evaluate. We go over all false positives per abstract marking ones that could be considered correct. Our evaluation shows that for FLAN-T5, 35 out of 73 entity and 12 out of 32 attribute errors are marked correct. For GPT4, these numbers are worse; 38 out of 126 entity and 20 out of 79 attribute errors are marked correct. This indicates that LLMs indeed struggle with our span extraction tasks, and their poor performance is not simply a consequence of strict evaluation.

6.2 How easily can we extend our schema to other domains?

Though we focus on extracting clinical findings from biomedical literature during schema design, we try to incorporate enough flexibility to allow our schema to be easily adapted to other scientific domains. To demonstrate this flexibilty, we conduct

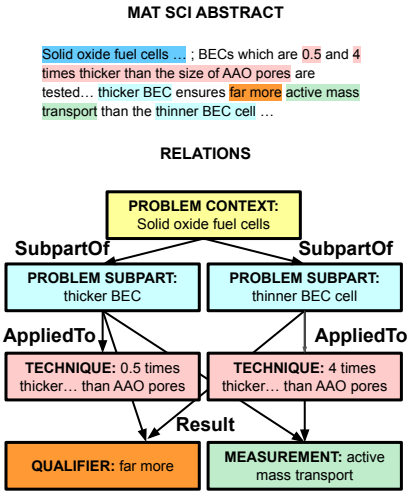

Figure 2: A partial example of entity, attribute and relation annotation using our generalized schema for a materials science abstract.

small-scale pilots in two additional domains: (i) Computer Science, and (ii) Materials Science.

We first develop a *generalized* version of our proposed schema for these studies. Of the three elements in our schema, entities and relations are largely transferable and only require minor renaming. Table 8 provides an overview of changes made to entity/relation nomenclature. Attributes on the other hand, were tailored more closely to our goal of extracting clinical findings. Therefore, we drop all attributes and ask our annotators to propose candidate attributes as they go through the annotation process. We use the same annotators who participated in dataset creation, to leverage their existing familiarity with our schema, assigning one annotator to each domain. Their task is to annotate ten abstracts each while documenting: (i) potential attributes that can be added to the schema, and (ii) important experimental information missed by the generalized schema.

After completing the task, annotators reported

that it was feasible to apply our proposed schemas to these scientific domains. Figure 2 shows an example materials science abstract with partial annotations according to the generalized schema. Computer science posed some difficulty due to the presence of lots of relative results and references in the abstract, which made entity annotation ambiguous. However, there were no important aspects of experimental information, aside from potential attribute proposals, that our current schema could not account for.

7 Conclusion

In this work, we presented CARE, a new IE dataset for the task of extracting clinical findings from biomedical literature. To collect this dataset, we first developed a new annotation schema capable of capturing fine-grained information about experimental findings, which unified several challenging IE phenomena such as discontinuous spans, nested relations and variable arity n-ary relations. Using this annotation scheme, we collected an extensively annotated dataset of 700 abstracts from clinical trials and case reports. Our benchmarking experiments showed that state-of-the-art extractive and generative LLMs including GPT4 still struggle on this task, particularly on relation extraction. We release both our annotation schema and CARE as a challenging new resource for the IE community and to encourage further research on extraction and representation of findings from scientific literature.

8 Limitations

Despite being a cornerstone of our work, the richness and complexity of our newly proposed annotation schema also poses some limitations. Annotators needed some prior experience with reading and understanding complex scientific text, and had to undergo multiple rounds of additional training before they were able to accurately apply our schema and start full-scale annotation. Though these stringent expertise and training requirements and heavy reliance on human annotators helped us collect a high-quality resource in CARE, they simultaneously limit the scalability of our collection protocol and make it difficult to construct large-scale benchmarks for this task, spanning multiple domains/fields of science.

Our annotated corpus, CARE, is based on RCTs and case reports. While our schema is broad and expressive enough to generalize to other experimental domains with minor adaptations, our generalization annotation studies were comparatively small and preliminary, limited to testing the schema on computer science and material science papers. In addition, while our schema covers many types of experimental findings, the richness and huge variety of scientific experiments neccessarily means that more types of findings could be added. In the future, more studies should be performed on using our schema in other domains, and on extending our schema with more types of information (entities, attributes, relations). CARE also focuses on English-language papers only, and in the future it would be interesting and important to extend our schema and dataset to cover biomedical/clinical studies in other languages, to capture important scientific findings that are potentially missed when only looking at papers in English.

Finally, a limitation of our current benchmarking effort is the lack of more flexible evaluation metrics, particularly when assessing the performance of generative LLMs. We try to provide supplementary human evaluation for some models to overcome this issue, but this is not scalable and would require ongoing/continuous evaluation efforts. This is not a major focus for our current work, but developing more flexible automated evaluation is an important future direction for IE research.

Acknowledgements

This work was funded in part by NSF Convergence Accelerator Award ITE-2132318. We would like to thank Lucy Lu Wang, Byron Wallace and Dr. Ruth Sapir-Pichhadze for valuable feedback on early versions of the annotation schema, and the Semantic Scholar team at AI2 and our anonymous reviewers for their helpful comments on our work.

References

Ziqi Chen, Bo Peng, Vassilis N Ioannidis, Mufei Li, George Karypis, and Xia Ning. 2022. A knowledge graph of clinical trials (ctkg). *Scientific reports*, 12(1):4724.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Vincent Claveau, Lucas Emanuel Silva Oliveira, Guillaume Bouzillé, Marc Cuggia, Claudia Maria Cabral Moro, and Natalia Grabar. 2017. Numerical

eligibility criteria in clinical protocols: annotation, automatic detection and interpretation. In *Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna, Austria, June 21-24, 2017, Proceedings 16*, pages 203–208. Springer.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An effective transition-based model for discontinuous NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5860–5870, Online. Association for Computational Linguistics.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MSˆ2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132.

Yanai Elazar and Yoav Goldberg. 2019. Where's my head? definition, data set, and models for numeric fused-head identification and resolution. *Transactions of the Association for Computational Linguistics*, 7:519–535.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213, Florence, Italy. Association for Computational Linguistics.

Rezarta Islamaj, Dongseop Kwon, Sun Kim, and Zhiyong Lu. 2020. Teamtat: a collaborative text annotation tool. *Nucleic acids research*, 48(W1):W5–W11.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Yanna Shen Kang and Mehmet Kayaalp. 2013. Extracting laboratory test information from biomedical text. *Journal of pathology informatics*, 4(1):23.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022a. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022b. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Aman Madaan, Ashish Mittal, Ganesh Ramakrishnan, Sunita Sarawagi, et al. 2016. Numerical relation extraction with minimal supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Ian H Magnusson and Scott E Friedman. 2021. Extracting fine-grained knowledge graphs of scientific claims: Dataset and transformer-based results. *arXiv preprint arXiv:2109.10453*.

Danielle L Mowery, Sumithra Velupillai, Brett R South, Lee Christensen, David Martinez, Liadh Kelly, Lorraine Goeuriot, Noemie Elhadad, Sameer Pradhan, Guergana Savova, et al. 2014. Task 2: Share/clef ehealth evaluation lab 2014. In *Proceedings of CLEF 2014*.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2022. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453, Seattle, United States. Association for Computational Linguistics.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207.

Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Lee M Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana K Savova. 2013. Task 1: Share/clef ehealth evaluation lab 2013. *CLEF (working notes)*, 1179.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

W Scott Richardson, Mark C Wilson, Jim Nishikawa, Robert S Hayward, et al. 1995. The well-built clinical question: a key to evidence-based decisions. *Acp j club*, 123(3):A12–A13.

Swarnadeep Saha, Harinder Pal, and Mausam. 2017. Bootstrapping for numerical open IE. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 317–323, Vancouver, Canada. Association for Computational Linguistics.

Olivia Sanchez-Graillet, Christian Witte, Frank Grimm, and Philipp Cimiano. 2022. An annotated corpus of clinical trial publications supporting schema-based relational information extraction. *Journal of Biomedical Semantics*, 13(1):1–18.

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and label: A two-stage identifier for nested named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2782–2794, Online. Association for Computational Linguistics.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*.

Asba Tasneem, Laura Aberle, Hari Ananth, Swati Chakraborty, Karen Chiswell, Brian J McCourt, and Ricardo Pietrobon. 2012. The database for aggregate analysis of clinicaltrials. gov (aact) and subsequent regrouping by clinical specialty. *PloS one*, 7(3):e33677.

Aryeh Tiktinsky, Vijay Viswanathan, Danna Niezni, Dana Meron Azagury, Yosi Shamay, Hillel Taub-Tabib, Tom Hope, and Yoav Goldberg. 2022. A dataset for n-ary relation extraction of drug combinations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3190–3203.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2023. Learning to generate novel scientific directions with contextualized literature-based discovery. *arXiv preprint arXiv:2305.14259*.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):52.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

A Schema Definitions

A.1 Entity Types

Entities can belong to one of the following seven types:

1. **Population:** Patient groups/cohorts studied in an article.

2. **Subpopulation:** Slices/sub-groups of a population entity sharing some underlying characteristic.
3. **Treatment:** Treatment regimens, procedures, therapies etc. prescribed and/or tested to alleviate a population's conditions/symptoms.
4. **Measurement:** Tests used to assess population status and outcomes of the tested intervention.
5. **Temporal:** Temporal information such as time points at which outcomes are measured.
6. **NumericFinding:** All numeric information associated with study findings (e.g., p-values, hazard ratios, etc.).
7. **Qualifier:** Non-numeric information associated with study findings that provides important perspective for interpreting them (e.g., phrases indicating evidence directionality).

A.2 Attribute Types

Attributes can belong to one of the following nine types:

1. **Age:** Numeric or non-numeric information about the age of the population under study.
2. **Sex:** Reported sex of the population under study.
3. **Size:** Size of the population sample under study.
4. **Condition:** Medical conditions prevalent in the study population, including diseases, symptoms, prior medical history and procedures, etc.
5. **Demographic:** Additional demographic information reported about the population such as location, race, etc.
6. **Route:** Description of the way an intervention is administered (e.g., a chemical may be administered orally, topically, intravenously, etc.).
7. **Dosage:** Quantity of administration for the intervention being studied. This is not necessarily limited to chemical/drug interventions (e.g., for an intervention like educational sessions, number of sessions is considered "dosage").
8. **Strength:** Strength of chemical/drug interventions administered.
9. **Duration:** Interval of time over which an intervention was administered.

A.3 Relation Types

Our schema allows for both binary and n-ary relations (with variable n), to capture four types of structure:

1. **AttributeOf:** N-ary relations linking population and intervention entities with their associated attributes.
2. **Subpopulation:** N-ary relations capturing parent-child relationships between population and subpopulation entities.
3. **InterventionOf:** Binary relations linking population and subpopulations entities with the intervention(s) tested on them.
4. **Result:** N-ary relations capturing all numeric or non-numeric outcome results and comparisons reported by linking together the population, subpopulation, intervention, measurement, numericfinding and/or qualifier and temporal entities involved in each result/comparison.

All n-ary relations can contain multiple entities of a single type. For example, a result relation can involve multiple interventions or populations. The only cardinality constraints imposed are that every result relation should focus on a *single* measurement entity and always contain *at least one* population/intervention entity.

B Additional Annotation Rules

While using this annotation schema to annotate clinical knowledge, we also keep in mind the following rules:

- For every entity/attribute span, only annotate its first occurrence in the text, unless there is a more descriptive span later. We follow this rule to avoid conducting an additional coreference annotation step to link all spans referring to the same entity.
- Ignore misspellings and include all associated modifiers and abbreviations while annotating spans
- Do not annotate generic or high-level spans (e.g., genetic disorder), or generic terms (e.g., complications, deficiency, disease, syndrome, gene, drug, protein, nucleotide, etc.).
- Do not annotate background occurrences of entities. For example, if a treatment Y is mentioned as "X is usually treated using Y,...", do not annotate Y unless Y was one of the treatments actually given to a population in the current study.

C Dataset Construction Details

Characteristics of sampled abstracts: Since the EBM-NLP corpus sampled randomized clinical trials from PubMed with an emphasis on cardiovascular diseases, cancer, and autism, the clinical trials portion of our dataset also heavily features these topics. On the other hand, for case reports, comparing MeSH term distributions across all reports (2M abstracts) with case reports containing

Round	Entity F1		Attribute F1		Relation F1	
	Exact	**Partial**	**Exact**	**Partial**	**Exact**	**Partial**
Pilot 1	0.6240	0.7579	0.7215	0.8163	0.2193	0.6379
Pilot 2	0.7206	0.8818	0.6923	0.7385	0.4997	0.7878
Pilot 3	0.6449	0.7900	0.5370	0.6852	0.4449	0.7960
Batch 1	0.5130	0.7318	0.7611	0.8496	0.3899	0.6979
Batch 2	0.6094	0.7900	0.6216	0.8508	0.6397	0.9137
Batch 3	0.5312	0.7797	0.6364	0.8182	0.3121	0.7595
Batch 4	0.5714	0.7817	0.7347	0.7755	0.5399	0.7343
Batch 5	0.5643	0.6929	0.4717	0.6762	0.3382	0.6766
Batch 6	0.6358	0.7930	0.5417	0.7582	0.3122	0.6890
Overall	**0.5764**	**0.7578**	**0.6174**	**0.7801**	**0.4209**	**0.7414**

Table 9: Evolution of inter-annotator agreement during pilots and full-scale annotation rounds

Type	Exact F1	Partial F1
Population	0.4333	0.8665
Subpopulation	0.4299	0.6168
Intervention	0.4333	0.5781
Measurement	0.5230	0.7554
Temporal	0.6230	0.6885
NumericFinding	0.7063	0.8812
Qualifier	0.6911	0.7749

Table 10: Inter-annotator agreement per entity type

Type	Exact F1	Partial F1
Age	0.8500	0.9756
Sex	0.9231	0.9231
Size	0.6462	0.7385
Condition	0.5091	0.7429
Demographic	0.6667	0.8000
Route	0.8000	0.8000
Dosage	0.6923	0.9630
Strength	-	-
Duration	0.0800	0.4800

Table 11: Inter-annotator agreement per attribute type. Note that the agreement sample did not include any strength entities.

Type	Exact F1	Partial F1
AttributeOf	0.7654	0.7654
InterventionOf	0.3797	0.3797
SubpopulationOf	0.1633	0.5185
Result	0.2561	0.7994

Table 12: Inter-annotator agreement per relation type

numeric information (the 900k we sample from), we see a massive reduction ($> 30\%$) in terms associated with the following topics: surgery and post-surgery care, dentistry, ophthalmology, prostheses and rehab, patient care and nursing, some mental disorders and circulatory diseases/issues. Hence, we expect these topics to be relatively undersampled in our pool of case reports.

Annotation Pilots: During pilots, we also conducted one or more disagreement discussion sessions per pilot round. These discussions were helpful in providing annotators the opportunity to highlight important spans/relations being missed by the schema, which led to the addition of the subpopulation entity, demographic attribute, and subpopulationof and treatmentof relations. Despite the introduction of some new elements, inter-annotator agreement continued to increase steadily over the pilot rounds, as shown in Table 9 before plateauing at the end of round 3.

Full-Scale Annotation: During rounds 1-3 of full-scale annotation, annotators were provided batches of 25 abstracts each. As their familiarity with the annotation schema and ability to handle ambiguous cases improved, we provided larger batches of 100 abstracts each during rounds 4-6. After each round, agreement was assessed and disagreement dicussions were conducted to discuss ambiguous cases, if needed, which ensured that agreement was maintained across rounds as seen from Table 9. Tables 10, 11 and 12 present final agreement scores per entity type, attribute type and relation type respectively. From these tables, we can see that Subpopulation and Intervention entities are the

trickiest to annotate, leading to lower agreement on SubpopulationOf and InterventionOf relation types due to error cascading (i.e., if entity annotations don't match, relation annotations are unlikely to match either).

D Inter-Annotator Agreement

Table 9 shows the evolution in inter-annotator agreement over our initial pilot rounds, as well as the level of inter-annotator agreement maintained during each round of the full-scale annotation process. We see a large increase in relation agreement from pilot 1 to pilot 2, and consistent agreement scores across all tasks in all rounds thereafter. Tables 10, 11 and 12 present inter-annotator agreement breakdown according to entity, attribute and relation types in our schema.

E Hyperparameter Details

Extractive Models:

- **OneIE:** We use an overall learning rate and weight decay of $1e-3$, and a learning rate and weight decay of $1e-5$ for the BERT component, a batch size of 10, and gradient clipping value of 5.0. The model is trained for 60 epochs with a 5-epoch warmup phase.
- **PURE:** We use a context window size of 300 words, overall learning rate of $1e-5$, task learning rate of $5e-4$, batch size of 16, and train for 100 epochs.
- **LocLabel:** We use a learning rate of $5e-6$, warmup rate of 0.1, weight decay of 0.01, gradient clipping value of 1.0, batch size of 6 and train for 35 epochs. LocLabel also requires word vectors, for which we use the 200-dimensional Pubmed-trained word2vec embeddings (BioWordVec) released by Zhang et al. (2019), which are available at `https://github.com/ncbi-nlp/BioWordVec`.
- **W2NER:** We use an overall learning rate of $1e-3$ and a learning rate of $5e-6$ for the BERT component, no weight decay, warmup factor fo 0.1, gradient clipping value of 5.0, batch size of 8, and train for 10 epochs.

Generative Models: All models are trained for 10 epochs with a learning rate of $1e-5$, input context length of 1024, output length of 128, and a batch size of 2.

GPT3.5/GPT4: We test the 16k and 8k context length versions of GPT3.5 and GPT4 respectively since our extraction tasks are abstract-level and require longer input contexts. We use the June 2023 versions of both models due to their *function calling* capabilities, which leverage a structured JSON output format to improve information extraction capabilities. All experiments are run with a temperature of 0 and max output length of 512 tokens.

F Computing Infrastructure

All LLM experiments are carried out on NVIDIA RTX A6000 GPUs with 48 GB RAM. Each finetuning run (FLAN-T5, BioGPT, BioMedLM) requires two GPUs with runtimes ranging from 9-17 hours depending on task size and model size. We use the DeepSpeed integration from Huggingface, with ZeRO-3 optimization, for multi-GPU training.

G Prompt Details

Figures 3, 4 and 5 present the prompts used to evaluate the performance of finetuned LLMs (FLAN-T5, BioGPT and BioMedLM) on entity, attribute and relation extraction respectively. Similarly, Figures 6, 7 and 8 present the prompts used to evaluate the performance of GPT-3.5 and GPT-4 models (in a zero-shot setting) on entity, attribute and relation extraction respectively. For the in-context learning setting, additional few-shot examples are appended to the prompt before providing the abstract.

Entity Extraction Prompt

Given an abstract from a biomedical paper, extract all population entities present in the abstract.
Population entities are defined as patient groups/cohorts studied in an article.
Generate the output in this format: entity1 <sep> entity2. If no entities of the specified type are present, output None.

Abstract: A primary intracranial HCG-producing tumour was studied in an 8 year old boy with congenital adrenal hyperplasia. The case provided a unique opportunity to study the sequential changes in serum and urinary androgens and HCG as measured by radioreceptor assay for HCG and by radioimmunoassay for HCG using antisera raised against the hormone specific for the beta subunit of HCG. Plasma concentrations of HCG, measured by the radioreceptor assay, closely correlated with the biologic activity of his tumour as measured by serum testosterone concentration. This case demonstrates that precocious puberty in any child, including with a known androgen disorder such as congenital adrenal hyperplasia warrants thorough investigation.

Extracted entities:

Output 🤖: "an 8 year old boy with congenital adrenal hyperplasia"

Figure 3: Example prompt used to evaluate the performance of finetuned LLMs on entity extraction. Such prompts are generated for all seven entity types in our dataset.

Attribute Extraction Prompt

Given an abstract from a biomedical paper with all population entities highlighted with **, extract all age attributes associated with these entities.
Age attributes are defined as numeric or non-numeric information about the age of the population under study.
Generate the output in this format: attribute1 <sep> attribute2. If no attributes of the specified type are present, output None.

Abstract: **A primary intracranial HCG-producing tumour was studied in an 8 year old boy with congenital adrenal hyperplasia**. The case provided a unique opportunity to study the sequential changes in serum and urinary androgens and HCG as measured by radioreceptor assay for HCG and by radioimmunoassay for HCG using antisera raised against the hormone specific for the beta subunit of HCG. Plasma concentrations of HCG, measured by the radioreceptor assay, closely correlated with the biologic activity of his tumour as measured by serum testosterone concentration. This case demonstrates that precocious puberty in any child, including with a known androgen disorder such as congenital adrenal hyperplasia warrants thorough investigation.

Extracted attributes:

Output 🤖: "8 year old"

Figure 4: Example prompt used to evaluate the performance of finetuned LLMs on attribute extraction with gold entities provided. Such prompts are generated for all nine attribute types in our dataset.

Relation Extraction Prompt

Given an abstract from a biomedical paper with all important entities and attributes highlighted with **, extract all attributeof relations between them.
Attributeof relations are defined as n-ary relations linking population and treatment entities with their corresponding attributes.
Generate the output in this format: entity1 <ent> attribute1 <sep> entity2 <ent> attribute2 <ent> attribute3. If no relations of the specified type are present, output None.

Abstract: **A primary intracranial HCG-producing tumour was studied in an 8 year old boy with congenital adrenal hyperplasia**. The case provided a unique opportunity to study the sequential changes in serum and urinary androgens and HCG as measured by radioreceptor assay for HCG and by radioimmunoassay for HCG using antisera raised against the hormone specific for the beta subunit of HCG. Plasma concentrations of HCG, measured by the radioreceptor assay, closely correlated with the biologic activity of his tumour as measured by serum testosterone concentration. This case demonstrates that precocious puberty in any child, including with a known androgen disorder such as congenital adrenal hyperplasia warrants thorough investigation.

Extracted relations:

Output 🤖: "a primary .. hyperplasia <ent> 8 year old <ent> boy <ent> a <ent> congenital adrenal hyperplasia"

Figure 5: Example prompt used to evaluate the performance of finetuned LLMs on relation extraction with gold entities and attributes provided. Such prompts are generated for all four relation types in our dataset.

Entity Extraction Prompt

System Prompt: Assistant is a system built to extract key information from biomedical papers.

User: Given an abstract from a biomedical paper, extract all entities present in the abstract.

Abstract: A primary intracranial HCG-producing tumour was studied in an 8 year old boy with congenital adrenal hyperplasia. The case provided a unique opportunity to study the sequential changes in serum and urinary androgens and HCG as measured by radioreceptor assay for HCG and by radioimmunoassay for HCG using antisera raised against the hormone specific for the beta subunit of HCG. Plasma concentrations of HCG, measured by the radioreceptor assay, closely correlated with the biologic activity of his tumour as measured by serum testosterone concentration. This case demonstrates that precocious puberty in any child, including with a known androgen disorder such as congenital adrenal hyperplasia warrants thorough investigation.

Output 🤖: {"population": ["an 8 year old boy with congenital adrenal hyperplasia"], "subpopulation": [], "treatment": [], "measurement": ["Plasma concentrations of HCG", "biologic activity of his tumour"], "temporal": [], "numericfinding": [], "qualifier": ["closely correlated with"]}

Figure 6: Prompt used to evaluate the performance of GPT-3.5 and GPT-4 on the entity extraction task. Additionally, entity type definitions from Table 1 are provided as *function parameters* to leverage OpenAI's function calling capabilities.

Attribute Extraction Prompt

System Prompt: Assistant is a system built to extract key information from biomedical papers.

User: Given an abstract from a biomedical paper with all the important entities highlighted with **, extract all attributes associated with these entities.

Abstract: **A primary intracranial HCG-producing tumour was studied in an 8 year old boy with congenital adrenal hyperplasia**. The case provided a unique opportunity to study the sequential **changes in** **serum** and **urinary** **androgens** and **HCG** as measured by radioreceptor assay for HCG and by radioimmunoassay for HCG using antisera raised against the hormone specific for the beta subunit of HCG. **Plasma concentrations of HCG**, measured by the radioreceptor assay, **closely correlated with** the **biologic activity of his tumour** as measured by serum testosterone concentration. This case demonstrates that precocious puberty in any child, including with a known androgen disorder such as congenital adrenal hyperplasia warrants thorough investigation.

Output 🤖: {"age": ["8 year old"], "sex": ["boy"], "size": ["A"], "condition": ["congenital adrenal hyperplasia"], "demographic": [], "route": [], "dosage": [], "strength": [], "duration": []}

Figure 7: Prompt used to evaluate performance of GPT-3.5 and GPT-4 on the attribute extraction task with gold entities provided. Additionally, attribute type definitions from Table 2 are provided as *function parameters* to leverage OpenAI's function calling capabilities.

Relation Extraction Prompt

System Prompt: Assistant is a system built to extract key information from biomedical papers.

User: Given an abstract from a biomedical paper with all the important entities and attributes highlighted with **, extract all relations between these entities and attributes.

Abstract: ** **A** primary intracranial HCG-producing tumour was studied in an **8 year old** **boy** with **congenital adrenal hyperplasia** **. The case provided a unique opportunity to study the sequential **changes in** **serum** and **urinary** **androgens** and **HCG** as measured by radioreceptor assay for HCG and by radioimmunoassay for HCG using antisera raised against the hormone specific for the beta subunit of HCG. **Plasma concentrations of HCG**, measured by the radioreceptor assay, **closely correlated with** the **biologic activity of his tumour** as measured by serum testosterone concentration. This case demonstrates that precocious puberty in any child, including with a known androgen disorder such as congenital adrenal hyperplasia warrants thorough investigation.

Output 🤖: {"subpopulationof": [], "treatmentof": [], "attributeof": [{"population": "A primary .. adrenal hyperplasia", "age": "8 year old", "sex": "boy", "size": "A"}], "result": [{"measurement": "plasma concentrations of HCG", "qualifier": "closely correlated with", "population": "A primary .. adrenal hyperplasia"}]}

Figure 8: Prompt used to evaluate performance of GPT-3.5 and GPT-4 on the relation extraction task with gold entities and attributes provided. Additionally, relation type definitions from Table 3 are provided as *function parameters* to leverage OpenAI's function calling capabilities.